A New Approach to Robust Speech Recognition Using FMQ, HMM and NN Based on A Robust LSP Distance Measure

Saf Asghar and Lin Cong
Advanced Micro Devices, Inc.
e-mail: saf.asghar@amd.com and lin.cong@amd.com

Abstract

In this paper a novel approach to robust speech recognition using Fuzzy Matrix Quantization (FMQ), Hidden Markov Model (HMM) and the Multi-layer Perception (MLP) Neural Network (NN) has been presented in the case of car noise environment with a proposed LSP distortion measure. The new isolated word speech recognition (IWSR) FMQ/HMM/MLP scheme has been discussed by combining the noise immunity learning process with interframe information results related to the envelopes and the probability-dependent maximum likelihood probability. Computer simulation results have clearly shown improved performance in recognition accuracy of the FMQ/HMM/MLP algorithm based on the robust LSP distance measure. The recognition accuracy approaches 95% with SNR at 10 dB.

1. Introduction

Next generation voice communication and information systems require efficient interaction mechanisms between users and terminals or remote database systems, therefore speaker dependent (SD)/independent (SI) isolated word speech recognition (IWSR) systems (algorithms) are being developed for this purpose. For example SI/SD, IWSR is an enabling technology for hands free dialing and interaction with voice store and forward systems, in mobile environments e.g., cars. Of course IWSR has received considerable attention in the last two decades but, there is still a challenge in designing robust IWSR systems capable of operating successfully at relatively low Signal to Noise Ratio (SNR) input conditions, especially when speech is corrupted by acoustic noise. The performance of existing medium to small vocabulary size IWSR schemes tends to deteriorate rapidly when the input SNR is below than 20 dBs.

Our previous work in robust IWSR systems excluded the use of acoustic noise reduction preprocessing and involved training the system using “clean” speech, during the IWSR design phase of the process, i.e., designing the systems in the mismatched noise condition (if the system is trained using the same type and level of noise that is expected to corrupt the input speech signal during recognition, which is called the matched noise condition) [1], [2]. Improved performance under noisy input conditions is mainly obtained by employing system components which are intrinsically robust enough to acoustic noise.

Within this general framework, there are two important factors which influence the robust operation of an IWSR scheme. These are: i) the representation of the speech short-term magnitude spectral envelope. The LSP coefficients are an example of such a robust representation and are used in our work, since band limited input distortion affects only a subset of the coefficients. ii) The “robust decision” reasoning employed on the spectral labelling/classification parts of the system. Soft decision algorithms offer advantages, when incorporated in the recognition process, as compared to equivalent schemes employing hard decision procedures. In our previous work, the principle of “soft decision” has been applied successfully in Vector Quantization (VQ) in the form of Fuzzy Vector Quantization (FVQ) [3]. FVQ is employed as a short - term spectrum labelling process and has been applied to IWSR systems, in conjunction with Hidden Markov Models (HMM) and Neural Networks (MLP). FVQ operates on single short-term spectra and, as a consequence, interframe information related to the “evolution” of the speech short-term spectral envelopes is not exploited by the IWSR system. This limitation can be overcome however with the introduction of Fuzzy Matrix Quantization, see[2].

This paper considers the case where IWSR system is designed and optimised, during training, using “clean” as well as “noise corrupted” speech signals in a very simple way. In particular, the new IWSR system employs a robust distance measure on MQ/FMQ spectral labelling followed by a Hidden Markov Model (MQ/HMM), or HMM and Neural Networks (MQ/HMM/MLP) classification techniques. The Fuzzy viterbi algorithm is employed in the HMM recognition process.

This robust performance ensures that a recognition accuracy of the order of 95% and 82% is obtained with input SNR values of 10 dB and 5dB respectively. The theory and structure of the proposed relatively small vocabulary, SD-IWSR schemes is presented in this paper together with recognition performance and computer simulation results based on extensive tests.

The paper is divided into six sections. Section 2 studies the presented hard matrix classification matrix which unifies the fuzzy clustering methodology. In section 3, the FMQ/HMM and FMQ/HMM/MLP systems are considered. The presented robust LSP distortion measure is discussed in section 4. Computer simulation results are given in section 5 and the conclusions are presented in section 6.

2. MQ and FMQ Classification Matrix

2.1 Conventional MQ Classification Matrix

Consider that a training set X TO speech spectral vectors, results in a set $X = \{x_1, x_2, \ldots, x_T\}$ of T, $P \times N$ matrices, where $T = \text{int}(TO / N)$

$$X_k = \begin{bmatrix} x_{11}^k & x_{12}^k & \ldots & x_{1N}^k & \\ x_{21}^k & x_{22}^k & \ldots & x_{2N}^k & \ldots \\ \ldots & \ldots & \ldots & \ldots & \\ x_{P1}^k & x_{P2}^k & \ldots & x_{PN}^k & \end{bmatrix} = \left[\overline{x}k(1), \quad \overline{x}k(2), \quad \ldots \quad \overline{x}k(N)\right]$$

where $\overline{x}k(j) = [x_{1j}^k \quad x_{2j}^k \quad \ldots \quad x_{Pj}^k]' \quad j = 1, 2, \ldots, N$, $k = 1, 2, \ldots, T$ is processed to yield a C-cell $A_i, i = 1, 2, \ldots, C$ partitioning of the Matrix space and a V-matrix entries MQ codebook containing C $v_i, \ i = 1, 2, \ldots, C, \ P \times N$, matrices

$$v_i = \begin{bmatrix} v_{11}^i & v_{12}^i & \ldots & v_{1N}^k \\ v_{21}^i & v_{22}^i & \ldots & v_{2N}^k \\ \ldots & \ldots & \ldots & \ldots \\ v_{P1}^i & v_{P2}^i & \ldots & v_{PN}^i \end{bmatrix} = [\overline{v_i}(1), \quad \overline{v}_i(2), \quad \ldots, \quad \overline{v}_i(N)]$$

where $\overline{v}_i(j) = [v_{1j}^i, \quad v_{2j}^i, \quad \ldots, \quad v_{Pj}^i]', \ j = 1, 2, \ldots, N$. The resulting quantization of X can be described by a $C \times T$ classification matrix U elements:

$$u_{ik} = \begin{cases} 0, & x_k \notin A_i \quad i = 1, 2, \ldots, C \\ 1, & x_k \in A_i \quad k = 1, 2, \ldots, T \end{cases}$$

Furthermore, the elements of this MQ matrix satisfy the two condition: $\sum_{i=1}^{C} u_{ik} = 1$ and $\sum_{k=1}^{T} u_{ik} > 0$. The columns O_j of the classification matrix U “map” effectively an input matrix x_j into a vector $O_j = \{u_{1j}, \quad u_{2j}, \quad \ldots, \quad u_{Cj}\}$ with all zero values except one element $u_{ij} = 1$ indicating that x_j to the i-th cell is performed so that the distortion

$$J(O_j, \ V) = \sum_{i=1}^{C} u_{ij} d(x_j, v_i) \qquad (2.1)$$

is minimized. $d(x_j, v_i)$ is distance measure

$$d(x_j, v_i) = \frac{1}{N} \sum_{n=1}^{N} d(\overline{x}_j(n), \overline{v}_i(n)) \qquad (2.2)$$

and in the research the distance measure $d(\overline{x}_j(n), \overline{v}_i(n)) = \sum_{m=1}^{P} (x_{mn}^j - v_{mn}^i)^2$ is the distance between the j-th column vector x_j and v_i , which is the centroid of the i-th cell. An optimum MQ codebook (partition of the matrix space) ensures that

$$J(U, \ V) = \sum_{j=1}^{T} \ \sum_{i=1}^{C} u_{ij} d(x_j, v_i) \qquad (2.3)$$

is minimized. Different distance measure derives different quantization for computing the “centroid” matrices v_i . This kind of matrix quantization classification is a conventional one, which is also defined as a hard decision.

2.2 Fuzzy Matrix Classification Matrix

Following similar arguments and definition as MQ discussed in section 2.1, the fuzzy matrix quantization [2] of **X** is described by a $C \times T$ fuzzy classification matrix U_F with elements $u_{ik} \in [0, 1]$, $i = 1, 2, \ldots, C, \quad k = 1, 2, \ldots, T$. The value of $u_{ik}, \ 0 \le u_{ik} \le 1$, indicates the degree of fuzziness of the k-th input matrix x_k to the i-th partitioning cell which is represented by the centroid v_i . The two conditions mentioned above are also satisfied [3]. In this case, u_{ik} is derived as:

$$u_{ik} = \frac{1}{\sum_{j=1}^{C}\left(\frac{d_{ik}(x_k, v_i)}{d_{jk}(x_k, v_j}\right)^{\frac{1}{(F-1)}}} \qquad (2.4)$$

where the constant F influences the degree of fuzziness. $d_{ik}(x_k, v_j)$ are the average distance measures as defined in section 2.1.

2.3 The Fuzzy Matrix Quantization Clustering Criterion

The columns O_j of the classification matrix U_F "map" an input matrix x_j into a vector $o_j = \{u_{1j,} \quad u_{2j,} \quad \ldots, \quad u_{Cj}\}$ results in the distortion

$$J(O_j,\ V) = \sum_{i=1}^{C} u_{ij}^{F} d(x_j, v_i) \qquad (2.5)$$

Furthermore, the overall distortion of the C entries fuzzy matrix quantizer operating on the **X** matrix set is

$$J(U,\ V) = \sum_{j=1}^{T} \ \sum_{i=1}^{C} u_{ij}^{F} d(x_j, v_i) \qquad (2.6)$$

Note that the summation of the O_j components is equal to unity. The largest component is the one which corresponds to the cell (centroid) with the smallest $d(x_j, v_i)$ value. O_j can be interpreted as a probability mass relating the input matrix x_j to all $v_i,\ i = 1, 2, \ldots, C$. Different distance measures derive different equations for computing the "centroid" matrices v_i [3].

Equation (2.3) and (2.6) provide the MQ and Fuzzy MQ distortion and can also be represented by the general distortion equation:

$$J(W, V) = \sum_{j=1}^{T} \sum_{i=1}^{C} w_{ij} d(x_j, v_i) \qquad (2.7)$$

where

$$w_{ij} = \begin{cases} u_{ij} & u_{ij} \in \{0, 1\} \\ u_{ij}^{F} & u_{ij} \in [0, 1] \end{cases}$$

In a same manner to FVQ, FMQ can be formed by Fuzzy C-means or by the Fuzzy LBG algorithms as discussed in [3], just extending that FMQ operates on N consecutive speech frames.

3. The Systems Description

The new FMQ/HMM/MLP system shown in figure 1, uses the FMQ as the front end of the HMM/NN classifier by adding the noise immunity into the training process of the codebook, HMMs and NN.

The speech and noise database consisted of NOISEX-92. The speech data consists of 10 English digits spoken by two speakers (one male and one female) with 40 utterances for each digit. 30 of the versions were used for training and the remaining were used for testing. Car noise is used as the noise input of the systems. A 10-th LSP analysis is performed every 20 msecs, allowing for a 10 msecs overlap between analysis frames.

Matrix Quantization is used as the front end of the HMM classifier, discussed in [2], since the feature matrices contain not only the short-term spectral feature, but also the temporal information embedded in the speech signal, so that a higher identification rate can be reached, compared with the Vector Quantization scheme.

During the training mode, the input for HMM classifier to learn is a combination of the output of the MQ part based on the clean speech and corrupted speech signals at different SNR levels. Fuzzy viterbi algorithm is used to produce the maximum likelihood probability obtained from HMM model based on word. The probability-dependent maximum likelihood probabilities with different SNR levels are used to train the post-classifier MLP. The characteristics of the proposed noise immunity system will be discussed in the following sections.

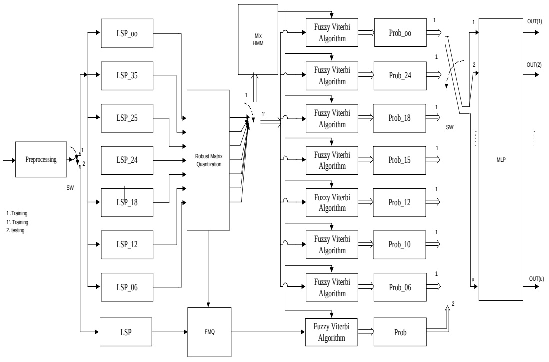

Figure 1. The FMQ/HMM/MLP system

MQ Design Process

The process of designing MQ involves the following three steps:

1. The training part of the input database is subdivided into nine sections D1 to D9 based on signals obtained at seven different car noise SNRs levels (∞dB, 35dB, 25dB, 24dB, 18dB, 15dB, 12dB, 10dB, 06dB). Each section consists of 300 words. Thus section D1 consists of 300 clean speech words,

section D2 consists of 300 speech words at 24 dB SNR and so on.

2. Each database section $D_i, i = 1, 2, \dots, 9$, provides and assembly LSP_i of vectors containing 10 LSP spectral coefficients and 210 assembly(database $D_i, i = 1, 2, 3, 4, 5, 7, 9$) are then mixed together to design one Matrix codebook for each word separately. The resulting Matrix codebook contains 160 codewords, 16 for each word. Notice that each entry is a P by N matrix [2].

HMM Design Process

The method described in [5] is used to set up an HMM $\lambda_j, j = 1, 2, ..., u$ (u is the size of the vocabulary) for each word vocabulary. However, in this case, the observation sequences $O = \{o_{1,} o_2, ..., o_{T_i}\}$ are now obtained from a given word at different car noise SNRs levels ($D_i, i = 1, 4, 5, 6, 7, 8, 9$), which are matrix quantised by the corresponding and previously designed robust codebook. Notice that we build separate HMMs for male and female.

MLP Training Process

The training database formed from 10 vocabulary words, each repeated 30 times for different input SNR conditions, can be presented as probability-dependent maximum likelihood probabilities:

Each database section $D_i, i = 1, 2, 4, 5, 7, 8, 9$, is used to generate a set of probability-dependent maximum likelihood probabilities. The MLP training process is organized so that a given version of a vocabulary word that has been produced at the k-th SNR input condition is computed by Fuzzy viterbi algorithm from a set of HMM $\lambda_j, j = 1, 2, ..., u$. This Fuzzy viterbi algorithm process generates a set of probability-dependent maximum likelihood probabilities which form the MLP input, see figure 1. Thus the MLP network is trained for the n-th vocabulary word, using the back propagation algorithm, by the seven SNR values in the same word version.

FMQ/HMM/MLP Recognition Process

When the system operates in a recognition mode, an input word W_j represented by a series $\{x_1, x_2, ..., x_{T_j}\}$ of T_j LSP version, is computed by the Fuzzy viterbi algorithm in parallel by u different HMMs. Thus, the probability-dependent, u-dimensional maximum likelihood probability vector

$$\overline{prob} = [prob_{1,} \quad prob_{2,} \quad \dots, \quad prob_u]$$

is presented to the MLP classification process: whose output $\{OUT(1), \quad OUT(2), \quad ..., \quad OUT(u)\}$, assume values in the region $0 \le OUT(j) \le 1$. The system classifies the input word W_j to the i-th vocabulary word if:

$$OUT(i) = \max\{OUT(1), \quad OUT(2), \quad ..., \quad OUT(u)\}$$

3.3 The MQ/HMM System Description

The FMQ/HMM/MLP improved recognition performance characters at low input SNR values can be attributed to the particular methodology used to expose the FMQ, HMM and MLP design processes to different input signal conditions. This powerful and general system training approach can be simplied as a further robust HMM based IWSR structure discussed in this section.

We can simply move the MLP post-classifier from the system FMQ/HMM/MLP shown in figure 1 to form the system MQ/HMM.

4. A Robust LSP Distance Measure

A new robust distance measure is proposed for the systems based on the LSP parameters discribed above, according to the information provided by speech signals corrupted by car noise. The study is useful for the robust speech recognition in noisy environment, in which the energy of noise is mainly located in low frequencies.

The most popular distance measure used for the systems based on LSP representations is the Euclidean measure. There are also some weighted LPC distance measures which have been presented [4].

We need to reword this by studying: by studying the effect of car noise on LSP speech parameters, we were able to determine which segments of the parameters work most affected:

$$d(f, \hat{f}) = \sum_{i=1}^{N_1} \alpha_1 [(f_i - e_i^{\beta_1} - \hat{f}_i)]^2 + \sum_{i=N_1+1}^{P} \alpha_2 [(f_i - \hat{f}_i) e_i^{\beta_2}]^2$$

where f_i and $\hat{f}_i$ are the i-th LSP in the test and reference vector respectively. e_i is the weight and frequency shift for the i-th LSP and is given by the LPC error power spectrum at the different test LSP frequencies. The constants α_1, α_2, β_1 and β_2 are experimentally determined. It is clear that when noise is large, the prediction error is large. When speech is corrupted by car noise, the frequency shift can compensate the car noise affect at low frequency part and this weighting can help at the high LSP frequency segment. In this paper, the α_1 is set to 1.6, α_2 is set to 0.68, β_1 is set to 0.5 and β_2 is set to 0.25.

5. Experiments and Results

The MQ/HMM system and FMQ/HMM/MLP system have been tested by using the NOISEX-92 - the test part of the database in computer simulation experiments. In these experiments, the matrix quantization length is chosen as N = 3. The systems are evaluated for recognition accuracy by presenting data with both of the same level SNRs used in the training data and a different SNR value which is not included in the training set.

MQ/HMM System Performance

Figure 2 shows MQ/HMM performances when using the proposed robust LSP distance measure (MQ/HMM_new) and the conventional LSP distance measure (MQ/HMM_old).

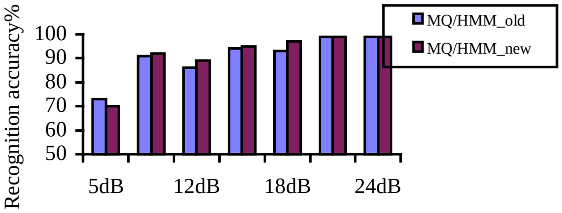

Figure 2: MQ/HMM system with different distance measures

It is shown that the MQ/HMM system based on the proposed the LSP distance measure the MQ/HMM system based on the conventional LSP distance measure. The recognition accuracy can be increased from 1% to 4% when SNRs below 20dB.

FMQ/HMM/MLP System Performance

In these experiments, the number of MLP hidden nodes P is 24. Figure 3 shows FMQ/HMM/MLP performances when using the proposed robust LSP distance measure (FMQ/HMM/MLP_new) and the conventional LSP distance measure (FMQ/HMM/MLP_old). The same conclusion can be obtained compared with the MQ/HMM_old system and the MQ/HMM_new system.

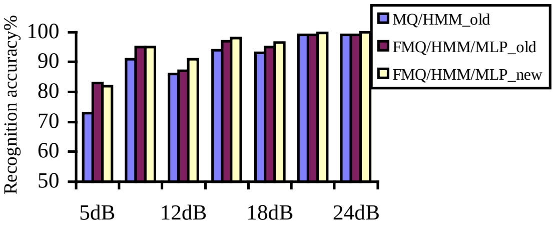

Figure 3: FMQ/HMM/MLP system with different distance measures

Also system FMQ/HMM/MLP and system MQ/HMM are compared in figure 4, in which we can see that the NN classifier trained by using the probability-dependent maximum likelihood probability can greatly increase the system’s performance especially in the case that the testing data is 5dB, which is not included in the training data SNR level. The recognition accuracy is 82% compared with the MQ/HMM system is 71% at 5dB. This shows that the method of gradually contaminating during training the input speech signal with noise, gives the MLP network a significant “noise immunity” capability. However, it is also shown that in the matched noise condition, we can use the simple MQ/HMM system from the simplicity point of implementing the SR system. For example, the recognition accuracy of the MQ/HMM system can reach to 92% when SNR is 10 dB.

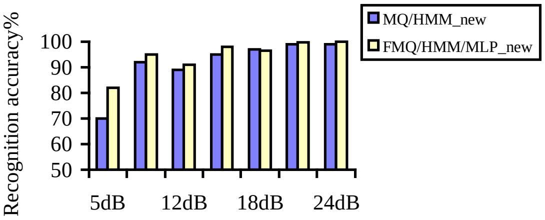

Figure 4. MQ/HMM and FMQ/HMM/MLP systems

6. Conclusions

This paper considers the case where an IWSR system is designed and optimised, during training, using clean as well as noise corrupted speech signals. In particular, two IWSR systems are proposed, which employ FMQ/MQ as the spectral labelling process, followed by a Hidden Markov Model (HMM), or a HMM and Neural Network (HMM/MLP) classification technique based on a new robust LSP distance measure. Both systems provide significant benefits in recognition accuracy, at low SNR input signal conditions by using a new and successful system training process and a new distance measure.

MQ/HMM achieve a recognition rate of 92% at 10 dB input SNR whereas at 20 dB SNR performance increases to 99%. The corresponding FMQ/HMM/MLP rates are 95% and 99%.

REFERENCES

[1] L. Cong, C. Xydeas and A. Erwood: ”Combining Fuzzy Vector Quantization and Neural Network Classification for Robust Isolated Word Speech Recognition”, ICCS’94, Vol.3, pp884-887, Nov., 1994, Singapore

[2] C. Xydeas and L. Cong: “Robust Speech Recognition in A Car Environment”, Intern. Conference on Digital Signal Processing, Vol. 1, pp84-89, June, 1995, Cyprus

[3] L. Cong, C. Xydeas and A. Erwood: ”A Study of Robust Isolated Word Recognition based on Fuzzy Methods”, EUSIPCO-94, Vol. 1, pp99-102, Sept., 1994, UK

[4] S. Furui and M. M. Sondhi: “Advances in Speech Signal Processing”, Marcel Dekker, Inc. 1992

[5] L. R. Rabiner: “A Tutorial on Hidden Markov Models and Selected Application Speech Recognition”, Proc. IEEE Vol. 77, pp257-286, 1989